

**Academic Performance Evaluation of Students (APES) :
Ubiquitous System, Analyzed:
Raw Scores to Letter Grades to GPA, is Unfair, by Design**

Dr. Keshava PRASAD Halemane,
Professor of Systems Analysis and Computer Applications,
Department of Mathematical and Computational Sciences,
National Institute of Technology Karnataka, Surathkal,
P. O. SrinivasNagar - 575025, Via: Mangalore, India.

kph@nitk.ac.in
kph_nitk@yahoo.com
kph_nitk@hotmail.com

Abstract:

A detailed and systematic analysis of the most imposing (impressive?!) and widely prevalent *Letter Grading System* (LGS) shows that it has some serious lacunae inherent in its very design, and fails to meet the essential purpose for which the Academic Performance Evaluation of Students (APES) is undertaken in schools/colleges/universities, as an integral component of the well accepted Teaching Learning Evaluation Review (TLER) Model Environment.

The LGS system behavior is shown to be highly chaotic. The system-intrinsic phenomena of Chaotic System Biased Information Loss/Corruption (CSBILC) is caused by the Quantization and the Contraction-Expansion Mapping that are implied in the LGS system computational model, used for the conversion of Raw Scores to Letter Grades and finally to GPA, CGPA, etc. Specifically, three kinds of chaotic system-intrinsic phenomena have been identified: Chaotic System Biased Amplification or Attenuation of Differentials (CSBAAD), Chaotic System Biased Suppression or Expression of Differentials (CSBSED), Chaotic System Biased Relative Rank Inversion (CSBRRI). Because of the poor design, although not deliberate, the LGS system is grossly unfair to the students class/community, who get subjected to Chaotic System Biased Unfair and Unreliable Comparisons (CSBUUC). It is unfair to the prospective recruiters or employers who expect some relevant, unbiased, reliable information to be conveyed through those transcripts or grade reports. It is unfair to the teachers, who find themselves utterly helpless, when the Raw Scores/Marks that they had originally assigned, are later subjected to chaotic system biased information loss/corruption. Also, the LGS system design presumes that the teacher's precision in evaluation is rather poor, limited to classification into possibly only a handful of distinct categories. However, there is an intrinsic contradiction in the system design philosophy itself since a significantly higher precision level is mysteriously presumed to have been achieved in reporting the final figures of GPA or CGPA, etc.

The LGS system just simply fails to provide a reliable mechanism for a true representation and communication of the appropriate/relevant information as to what the teachers really meant to convey, regarding the measurements originally conducted, towards the academic performance evaluation of students (APES), that is (should be) unquestionably considered an integral component of the overall *Teaching Learning Evaluation Review* (TLER) *Environment*. As an alternative, the "*Students Academic Performance Evaluation System* (SAPES)" is proposed.

Index Terms:

Raw-Scores, Letter-Grades, Grade-Point-Average (GPA),
Chaotic System Biased Information Loss/Corruption (CSBILC),
Chaotic System Biased Unfair and Unreliable Comparisons (CSBUUC).

1. Introduction

The most widely prevalent *Letter Grading System* (LGS) model requires each of the teachers as evaluators, to assign an appropriately chosen letter grade, from among the set of assignable letter grades defined by the school/college/university, to each of the students in each of the courses/subjects/papers at the end of every academic session/term/quarter/semester. For this purpose, the teacher may conduct a sequence of tests/examinations/etc, (an almost continuous, cumulative/incremental evaluation) that together facilitate in measuring the extent/degree of proficiency achieved by the student in the subject, based on the performance of the student in such a sequence of tests etc, which when well designed would evenly cover the entire subject material that a student is expected to master during that period of study.

This paper presents a systematic analysis of the process/mechanism designed and incorporated into the existing LGS system for the Academic Performance Evaluation of Students (APES) and shows that there are some serious inherent lacunae in the very design of the LGS system. In particular, it is shown that the system exhibits a chaotic, highly complex, counter-intuitive as well as undesirable behavior. The system-intrinsic phenomena of Chaotic System Biased Information Loss/Corruption (CSBILC) is caused by the Quantization and the Contraction-Expansion Mappings that are implied in the LGS system computational model, used for the conversion of Raw Scores to Letter Grades and finally to GPA, CGPA, etc. This results in chaotic system biased unfair and unreliable comparisons (CSBUUC), among the students being evaluated, thus making it to be very grossly unfair to the students community. It is also unfair to the prospective recruiters or employers who expect some relevant, unbiased, reliable information to be conveyed through those transcripts or grade reports. Again, it is very unfair to the teachers, who find themselves utterly helpless, when the Raw Scores/Marks that they had originally assigned, are later subjected to chaotic system biased information loss/corruption.

Also, the LGS system design presumes that the teacher's precision in evaluation is rather poor, limited to classification into possibly only a handful of distinct categories. However, there is an intrinsic contradiction in the system design philosophy itself since a significantly higher precision level is mysteriously presumed to have been achieved in reporting the final figures of GPA or CGPA, etc.

The LGS system just simply fails to provide a reliable/precise and robust/resilient mechanism for a true representation and communication of the appropriate/relevant information as to what the teachers really meant to convey, regarding the unbiased/objective and fair/justifiable measurements originally conducted, towards the academic performance evaluation of students

(APES), that is (should be) unquestionably considered an integral component of the overall *Teaching Learning Evaluation Review (TLER) Model Environment*. As an alternative, "Students Academic Performance Evaluation System (SAPES)" is proposed.

2. A Typical Letter Grading System (LGS) Model

Although there are several minor variations among the different grading systems that are prevalent in various schools, in terms of the various parameters incorporated in such a system design, the essential system behavior observed in these different scenarios happen to be the same! The observed differences and variations among them are only in terms of the nature and/or the extent and/or the locus, of the very same, essentially general, system behavior, which we analyze here, by taking a typical grading system model, although the same observations/comments can be made with regard to any other specific grading system model, belonging to the wide spectrum of various possible distinctly different specific system models.

Let us suppose that the set of valid assignable *Letter Grades* and the associated *Quality Points* be as follows:

A = 5.00, B = 4.00, C = 3.00, D = 2.00, E = 1.00, F(fail) = 0.00

Other letter grades (like Pass/Fail, etc) may not carry *Quality Points*, but only indicate certain essential information about the Academic Progress of a student towards one's Graduation Requirements. Once a letter grade is assigned to each of the courses that a student had "*registered for credit*", the *Grade Point Average (GPA)* for that student is computed as the weighted average of the *Quality Points* associated with the corresponding letter grades, with the *Course Credits/Units (CCU)* of the corresponding courses as the multiplicative weighting factors for this computation.

Now let us focus on the problem that a typical teacher would have to face, in order to assign an appropriate letter grade to a student on a particular subject. The teacher does go through the evaluation of the answer scripts/papers of a sequence of tests, in order to determine the final letter grade based on some well defined mechanism usually announced to the students at the very beginning of the academic term. For example, a teacher may decide and therefore announce to the students that there would be four one-hour tests each carrying 25% weightage in determining the final grade. Now, how is a test paper evaluated? How, in a given test paper, the answer to each of the questions, evaluated? And how are these combined together (aggregation process/mechanism) to form what we call as the Raw-Score for a test paper?

We believe, that no teacher starts off with assigning appropriately chosen letter grades (from among the set of assignable letter grades as defined by the school) to each of the various questions in a test. Usually, while setting the Question Paper for a test, appropriate numerical marks/scores are assigned for each of the questions, maybe on a Numerical Scale in the range 0-100 (Percentage Scale). While evaluating a test answer paper, appropriate mark/score is given for the answer to each of the questions, based on how close it is to the expected answer, or to what extent the answer is acceptable (in case of possible multiple correct/acceptable answers, like in design problems), and these numerical scores are simply

added together to come up with a total numerical score, which is the Raw-Score for that particular test paper.

If in fact it be possible, to assign letter grades right at the level of evaluation of each of the questions in a test, and if in fact there is a well accepted standard procedure designed to somehow combine them into an aggregate letter grade for that particular test paper, and then again combine such letter grades for each of the four or more tests into a final letter grade for that student in that specific course; then I do not have much to say in this paper! The whole problem arises because one needs to resort to numbers, in order to have a well defined rational scheme for measurements, and also to enable us in combining these various scores into a consolidated overall measure of whichever entity is being measured.

Even assuming that each teacher uses some mapping scheme to convert the raw scores to the letter grades, usually the nature of such a mapping scheme may be of two generic types (although there can be any number of specific mapping schemes): One of them is called '*grading on the curve*', or '*distribution based evaluation*'; wherein the letter grades are force-fitted onto a standard normalized distribution curve, like the Gaussian Probability Distribution Curve. Since the actual distribution of the raw scores in a class is somewhat independent of the teacher and the system, the only way that the letter grades can be made to fit into a given distribution is to have the mapping scheme dependent on the actual distribution of the raw scores in the class; hence this particular name for it. This approach has several proponents and/or followers. However, I believe that *any force-fitting performed on any set of raw data leads to Information Loss/Corruption, and "we can only end up seeing what we look for"*. Therefore, for the purpose of having a typical grading system model, I shall not use this scheme. This decision in itself, is not a limitation on the scope of validity or the conclusions that can be drawn from this analysis (being seemingly dependent on the system model) since any mapping scheme, belonging to any of these two or any other type, generic or otherwise, would have the *same common characteristics* which in fact cause the system to exhibit the undesirable behavior that is explained in this paper.

Another mapping scheme uses some fixed well defined divisions of the entire scale (raw scores scale) into intervals, which I refer to later in this paper as the "*Quantization Interval Domains*". Each such interval is mapped onto a single point in the quality point scale. The definition of these quantization interval domains, and also the associated mapping scheme is independent of the actual distribution of the raw scores in a class. An appropriate term to describe this approach would be "*distribution independent grading scheme*"; to indicate that the mapping scheme used for converting a raw score to a letter grade is independent of the actual distribution of the raw scores in a class; and also that the resulting letter grades are not force-fitted into any specific distribution, standard, normalized or otherwise!

So, every teacher does the evaluation of the test papers based on a numerical scale, and assigns a numerical score (say for example, in the range 0-100, as in a percentage scale) as the raw score for each of the test papers. These raw scores can therefore be combined together to determine an overall raw score for a student in a course at the end of an academic term, like, the weighted average of the four test scores, with equal weights if so announced earlier, or any other way. Now, in order for the teacher to be able to assign an appropriately chosen letter grade to a student in a

subject, at the end of the academic term, there has to be a well defined (specified by or accepted by the school) or at least some rational method of mapping the raw scores to the letter grades. In the absence of such a mapping scheme, each teacher may be compelled to design one, whenever the need arises (and such ad hoc nature, in itself, would be an even more serious lacuna in the system). For the purpose of our model, let us assume that an accepted mapping scheme has been made available to the teacher, possibly evolved over a period of time, among the faculty, and we shall take it to be as follows:

100-90 : A, 89-80 : B, 79-70 : C,
69-60 : D, 59-50 : E, 49-00 : F(fail)

This mapping scheme, to convert the *Numerical Raw Scores* to *Letter Grades*, along with the above defined (specified by the school) *Quality Points* for each of the letter grades; can be seen to result in an overall computational procedure for our typical system model, in order to determine the letter grade for each of the courses, for each of the students, using the available numerical raw scores; and also to compute the *Grade Point Average (GPA)*, at the end of an academic term; and similarly the *Cumulative Grade Point Average (CGPA)* for the programme of study.

3. Measuring the Perimeter of a Pentagon

In order to start an analysis of the above described typical Letter Grading System (LGS) model, let us first pose a simple problem to ourselves and try to find an acceptable solution. Let us look at a scenario wherein you are required to measure the length of the perimeter of a pentagon, and report the result in millimeters (p mm), and then you are required to compute the average length of a side of that pentagon and report that also in millimeters (s mm). Suppose I offer to help you perform this task, but with a small difference; first I measure the length of each of the five sides of the pentagon in millimeters, then round-off each of these to the nearest centimeter, add them up, and convert the sum back to millimeters, and report the result (r mm), and using this I compute the average length of a side in millimeters (t mm). Would you accept my offer?

Let us say, for the sake of concreteness, that there are two pentagons P1, P2, with the actual lengths of their sides (in mm) being as follows:

P1: {78, 77, 76, 72, 71}, thus having its perimeter, p1 = 374 mm, and

P2: {79, 78, 74, 73, 72}, thus having its perimeter, p2 = 376 mm.

If I were to round-off the length measurement of every side to nearest centimeter, the results would be:

P1: {8, 8, 8, 7, 7}, perimeter = 38cm, that is, r1 = 380 mm,

P2: {8, 8, 7, 7, 7}, perimeter = 37cm, that is, r2 = 370 mm.

The average length of a side, for these two pentagons is as follows:

P1: the average length of a side is, s1 = 74.8 mm, and

P2: the average length of a side is, s2 = 75.2 mm.

If I were to use the result of my measurements (with term-wise round-off)

the average length of the sides, for these two pentagons would be as follows:

P1: the average side length would be, t1 = 76 mm, and

P2: the average side length would be, t2 = 74 mm.

Now, we notice that ($p_1 < p_2$) but ($r_1 > r_2$), or equivalently, that ($s_1 < s_2$) but ($t_1 > t_2$). In fact, in this example, it turns out that ($p_2 - p_1$) is just 2mm, whereas ($r_2 - r_1$) is -10mm, which means that there is not only a *magnification of the differences*, but also a *sign reversal*.

The results associated with the above round-off procedure are certainly not acceptable. It is to be noted here that the procedure suffers from *Information Loss/Corruption*; because of *Round-off Errors*. Such a system of measurement and computation, exhibits a chaotic, highly complex, counter-intuitive as well as undesirable system behavior, due to the Sharp Interval Domain Boundaries across which distinctly different decisions are followed regarding the round-off procedure.

One may remind oneself of the well established standard procedure usually adopted in numerical computations to use double precision arithmetic for all intermediate computations, although the input as well as the expected output happen to be in single precision representations. The double precision representation of all the intermediate computations allows sufficient room to accommodate for the cumulative errors of computations (arising due to the finite word size arithmetic, etc.) so that at the end the result would still be acceptable at a single precision accuracy level. On the other hand, imagine (as is well known) what would be the consequences (in terms of the errors involved) if the input and output were required to be in double precision, whereas all the intermediate computations were to be performed in single precision arithmetic.

What does this scenario have to do with the Academic Performance Evaluation of Students (APES), using the most widely prevalent Grading System Policies and Procedures; namely the conversion of the raw scores to letter grades, and then to the Quality Points, and finally the computation of the Grade Point Average (GPA) as the weighted average of these quality points? Yes, indeed there is more than just a semblance of an analogy here! The sides of the pentagon could be the five subjects, then their lengths would be the raw scores (percentage; or letter grade) the perimeter would be the total score (p , the total raw score; or r , the sum total of the product of the quality points for each subjects multiplied by the corresponding Course Credits/Units (CCU), as the numerator computed in the GPA calculation); and the average side length computed would be the overall average score (s , the overall average score computed directly as a weighted average of the raw scores; or t the usual Grade Point Average or GPA).

One may wonder whether the situation would be better in the above example problem, if we were to use a different round-off procedure, like rounding downward to the Floor, or rounding upward to the Ceiling, instead of the standard round-off procedure. A further look into the details of the system behavior shows without any ambiguities, that such a system behavior arises not particularly from the specific mechanism or procedure for the round-off, but rather from the fact that there has been a round-off procedure applied in the measurement and computational system. For example, consider $P_1: \{83, 82, 81, 75, 74\}$ and $P_2: \{85, 81, 79, 78, 77\}$, that is with $s_1=79, s_2=80$; giving rise to $t_1=76, t_2=74$ with downward round-off procedure, whereas $t_1=86, t_2=84$ with upward round-off procedure; giving rise to a magnification and sign reversal, between the actual differences and the observed differences (as per any of the round-off procedures).

4. Chaotic LGS System Behavior

Consider a class of 24 students, identified as S01, S02, . . . S24, each taking five courses, of which one (s1x4c) is of 4 CCU, three (s2x3c, s3x3c, s4x3c) are of 3 CCU, and one (s5x2c) is of 2 CCU. Let us suppose that at the end of their academic term, they get the following grades:

one of the students, S01, gets 5 A's and 0 B's,
 each of the 5 students, S02-S06, get 4 A's and 1 B's,
 each of the 6 students, S07-S12, get 3 A's and 2 B's,
 each of the 6 students, S13-S18, get 2 A's and 3 B's,
 each of the 5 students, S19-S23, get 1 A's and 4 B's,
 one of the students, S24, gets 0 A's and 5 B's,

as given in the accompanying Table. Now, suppose that the actual Raw-Scores that these students earned in their courses, were to be as indicated in the Table below, which also includes the complete set of information on the scores obtained by these students.

Student I.D.	Raw-Score (Letter-Grade) subject-wise					SAPI%	SAPI-	GPA	GPA-
	s1x4c	s2x3c	s3x3c	s4x3c	s5x2c	(sRSCCU)	Rank	(sQPCCU)	Rank
S01	90 (A)	90 (A)	90 (A)	90 (A)	90 (A)	90 (1350)	8-12	5.00 (75)	1
S02	89 (B)	99 (A)	99 (A)	98 (A)	98 (A)	96 (1440)	1	4.73 (71)	4-6
S03	90 (A)	93 (A)	89 (B)	93 (A)	90 (A)	91 (1365)	4-7	4.80 (72)	3
S04	89 (B)	90 (A)	90 (A)	90 (A)	92 (A)	90 (1350)	8-12	4.73 (71)	4-6
S05	90 (A)	90 (A)	91 (A)	90 (A)	81 (B)	89 (1335)	13-17	4.87 (73)	2
S06	81 (B)	91 (A)	90 (A)	91 (A)	90 (A)	88 (1320)	18-21	4.73 (71)	4-6
S07	98 (A)	98 (A)	98 (A)	89 (B)	89 (B)	95 (1425)	2	4.67 (70)	7-8
S08	89 (B)	90 (A)	97 (A)	90 (A)	89 (B)	91 (1365)	4-7	4.60 (69)	9-10
S09	89 (B)	89 (B)	90 (A)	91 (A)	92 (A)	90 (1350)	8-12	4.53 (68)	11-12
S10	90 (A)	91 (A)	91 (A)	89 (B)	81 (B)	89 (1335)	13-17	4.67 (70)	7-8
S11	85 (B)	90 (A)	90 (A)	90 (A)	85 (B)	88 (1320)	18-21	4.60 (69)	9-10
S12	81 (B)	81 (B)	91 (A)	90 (A)	90 (A)	86 (1290)	22	4.53 (68)	11-12
S13	98 (A)	98 (A)	88 (B)	89 (B)	89 (B)	93 (1395)	3	4.47 (67)	13-14
S14	95 (A)	88 (B)	89 (B)	88 (B)	95 (A)	91 (1365)	4-7	4.40 (66)	15-16
S15	89 (B)	89 (B)	89 (B)	92 (A)	92 (A)	90 (1350)	8-12	4.33 (65)	17-18
S16	90 (A)	91 (A)	89 (B)	89 (B)	84 (B)	89 (1335)	13-17	4.47 (67)	13-14
S17	90 (A)	87 (B)	86 (B)	87 (B)	90 (A)	88 (1320)	18-21	4.40 (66)	15-16
S18	81 (B)	81 (B)	81 (B)	90 (A)	90 (A)	84 (1260)	23	4.33 (65)	17-18
S19	98 (A)	88 (B)	89 (B)	88 (B)	89 (B)	91 (1365)	4-7	4.27 (64)	19-21
S20	89 (B)	89 (B)	89 (B)	88 (B)	98 (A)	90 (1350)	8-12	4.13 (62)	23
S21	90 (A)	89 (B)	89 (B)	89 (B)	87 (B)	89 (1335)	13-17	4.27 (64)	19-21
S22	88 (B)	87 (B)	90 (A)	87 (B)	88 (B)	88 (1320)	18-21	4.20 (63)	22
S23	90 (A)	80 (B)	80 (B)	81 (B)	81 (B)	83 (1245)	24	4.27 (64)	19-21
S24	89 (B)	89 (B)	89 (B)	89 (B)	89 (B)	89 (1335)	13-17	4.00 (60)	24

The Students Academic Performance Index (SAPI) is computed as follows: First, the sum (sRSCCU) of the raw scores, each of them being multiplied by the corresponding Course Credits/Units, is computed. Then, sRSCCU is divided by the sum (sCCU) of Course Credits/Units (15, here) registered by the student for the semester; the ratio is the required SAPI value. To compute the GPA, first the sum (sQPCCU) of the Quality Points (associated with the Letter Grades) each of them being multiplied by the corresponding Course Credits/Units, is computed. Then, sQPCCU is divided by the sum (sCCU) of Course Credits/Units (15, here) registered by the student for the semester; the ratio is the required GPA value.

The accompanying Table gives, for each student in the class, the Raw Score and the Letter Grade for each of the five courses, SAPI% (SAPI percentage), sRSCCU, SAPI-Rank, GPA (out of 5.00), sQPCCU, GPA-Rank. The computations are as explained above.

The above computations do reveal some bare naked truth, about the lacunae or weaknesses or drawbacks in the prevailing grading system design. Specifically, the following observations are made.

Sometimes small difference in SAPI is associated with large difference in GPA:

(S02-S07-S13-S19-S24), (S06-S12-S18-S23); and

sometimes large difference in SAPI is associated with small difference in GPA:

(S03-S06), (S07-S09), (S10-S12), (S07-S12),
(S13-S15), (S16-S18), (S13-S18), (S19-S22);

which I call as the *Phenomenon of Chaotic System Biased Amplification or Attenuation of Differentials (CSBAAD)*.

Also, sometimes large difference in SAPI is associated with no difference in GPA:

(S02-S04-S06), (S07-S10), (S08-S11), (S09-S12),
(S13-S16), (S14-S17), (S15-S18), (S19-S21-S23); and

sometimes no difference in SAPI is associated with large difference in GPA:

(S01-S04-S09-S15-S20),
(S03-S08-S14-S19), (S04-S09-S15-S20),
(S05-S10-S16-S21-S24), (S06-S11-S17-S22);

which I call as the *Phenomenon of Chaotic System Biased Suppression or Expression of Differentials (CSBSED)*.

Again, we also observe that sometimes an increase in SAPI may result in a decrease in GPA and vice versa, as can be observed in cases:

(S02-S03), (S02-S05), (S03-S05), (S04-S05),
(S08-S10), (S09-S10), (S09-S11), (S14-S16), (S15-S16), (S15-S17),
(S20-S21), (S20-S22), (S20-S23), (S22-S23); and

the more drastic and shocking observation, that sometimes -

4 A's and 1 B's, should have been better than all 5 A's,
(S02 > S01), (S03 > S01);

3 A's and 2 B's, should have been better than all 5 A's,
(S07 > S01), (S08 > S01);

2 A's and 3 B's, should have been better than all 5 A's,
(S13 > S01), (S14 > S01);

1 A's and 4 B's, should have been better than all 5 A's,
(S19 > S01);

3 A's and 2 B's, should have been better than 4 A's and 1 B's,
(S07 > S03), (S07 > S04), (S08 > S04),
(S07 > S05), (S08 > S05), (S09 > S05),
(S07 > S06), (S08 > S06), (S09 > S06), (S10 > S06);

2 A's and 3 B's, should have been better than 4 A's and 1 B's,
(S13 > S03), (S13 > S04), (S14 > S04),
(S13 > S05), (S14 > S05), (S15 > S05),
(S13 > S06), (S14 > S06), (S15 > S06), (S16 > S06);

1 A's and 4 B's, should have been better than 4 A's and 1 B's,
(S19 > S04), (S19 > S05), (S20 > S05),
(S19 > S06), (S20 > S06), (S21 > S06);

just 5 B's only, should have been better than 4 A's and 1 B's,
(S24 > S06);

2 A's and 3 B's, should have been better than 3 A's and 2 B's,
(S13 > S08), (S13 > S09), (S14 > S09),
(S13 > S10), (S14 > S10), (S15 > S10),
(S13 > S11), (S14 > S11), (S15 > S11), (S16 > S11),
(S13 > S12), (S14 > S12), (S15 > S12),
(S16 > S12), (S17 > S12);

1 A's and 4 B's, should have been better than 3 A's and 2 B's,
(S19 > S09), (S19 > S10), (S20 > S10),
(S19 > S11), (S20 > S11), (S21 > S11),
(S19 > S12), (S20 > S12), (S21 > S12), (S22 > S12);

just 5 B's only, should have been better than 3 A's and 2 B's,
(S24 > S11), (S24 > S12);

1 A's and 4 B's, should have been better than 2 A's and 3 B's,
(S19 > S15), (S19 > S16), (S20 > S16),
(S19 > S17), (S20 > S17), (S21 > S17),
(S19 > S18), (S20 > S18), (S21 > S18), (S22 > S18);

just 5 B's only, should have been better than 2 A's and 3 B's,
(S24 > S17), (S24 > S18);

just 5 B's only, should have been better than 1 A's and 4 B's,
(S24 > S22), (S24 > S23);

which I call as the *Phenomenon of Chaotic System Biased Relative Rank Inversion (CSBRRRI)*.

It is to be noted here that the above observed system behavior is certainly not because of the specific choice of the set of various parameters and their values, but in fact an intrinsic characteristic of a poor system design, which happen to be exposed very well through this example scenario. This chaotic system behavior is neither arising from nor is avoidable by any teachers through their decisions on the evaluation itself. Also, such system behavior is neither associated with nor can be avoided by any general or specific students' action through their academic performance levels, whether individually or otherwise. The various parameters, like the number of courses considered for the analysis, the Course Credits/Units (CCU), the

number of students being considered, the relative performance levels among them, the actual distribution of the raw scores in a class, and even the actual grading system model (policies and procedures) adopted, can in fact possibly be changed within their corresponding ranges of usually acceptable variabilities, and still we can construct any number of example scenarios like the above, to illustrate such system behavior.

Even for the sake of simplicity, if we were to consider a single course, for our analysis; all students with raw scores anywhere in the range from 90 to 99+ (interval size being 10) get the grade A, no distinctions being made among them! But a student who gets a raw score of 89 is put in the grade B category, as distinct from a student getting a raw score of 90 who is put in the grade A category; thus leading to a situation where a very small difference of just 01% in the raw-score gets blown up by a magnification resulting in a difference of 20% in the Quality Points gained on that course. The extent of this undesirable magnification or amplification (or sometimes, even attenuation, as seen in the above explained example scenario) of the differences, is in fact dependent upon the relative location of the various raw-scores within the corresponding *Quantization Interval Domain* (ref: next section), and also their proximity to the *Quantization Interval Domain Boundaries*.

What do we conclude from the above observations? We all agree that every teacher really means whatever s/he initially gives as the Raw Score for each of the test papers. All other transformations, conversions or computations, are only performed later, starting with these raw scores as the original raw input data; in order just to conform to the Policies and Procedures of the prevailing LGS System as per the Rules and Regulations in the school. However, the "Student Academic Performance Index (SAPI)" computed directly as a weighted average of the *original Raw Scores*, is certainly the best and in fact a true representation, of whatever the teachers really intended while originally evaluating each of their students.

Unfortunately, the prevailing LGS System just does not provide a reliable mechanism for a teacher to convey the information regarding the student academic performance evaluation, through the transcripts/grade-reports, to whoever is really concerned in receiving (and possibly even acting, based on) such information! The system behavior exhibiting the *Phenomenon of Chaotic System Biased Amplification or Attenuation of Differentials (CSBAAD)*, the *Phenomenon of Chaotic System Biased Suppression or Expression of Differentials (CSBSED)*, and the *Phenomenon of Chaotic System Biased Relative Rank Inversion (CSBRRRI)*, are certainly extremely undesirable, or rather just simply *unacceptable*.

The system is grossly *unfair to the students*, because they get subjected to *chaotic system biased unfair and unreliable comparisons (CSBUUC)*, of course, not caused by any individual person, but because of the poor system design, which again is not a result of any deliberate intent of anyone. The system is *unfair to the recruiters* or prospective employers, because they are *misled by unreliable and/or misleading information*; when in fact, they deserve to be provided with the best and most reliable information on the students academic performance; by unbiased and fair measurements, with appropriately designed reliable indicators, and communicated through the transcripts or grade-reports. The system is *unfair to the teachers* who although *helpless* (because of the existing system imposed on them, to be followed) are looked upon as the possible perpetrators of such a disturbingly chaotic system, which is both unfair as well as unreliable; wherein the Raw Scores/Marks that they had

originally assigned, are later subjected to chaotic system biased information loss/corruption. Also, the LGS system design presumes that the teacher's precision in evaluation is rather poor, limited to classification into possibly a handful of distinct categories. However, there is an intrinsic contradiction in the system design philosophy itself since a significantly higher precision level is mysteriously presumed to have been achieved in the final figures of GPA or CGPA, etc.

The prevailing LGS System just simply fails in meeting the most essential and basic necessary condition of meeting the Fundamental Quality Characteristics of being *unbiased/objective and fair/justifiable*; in providing a *reliable/precise and robust/resilient* mechanism for the representation and the communication of the *appropriate/relevant* information about what the teachers or evaluators really meant in their evaluations; for which the whole exercise of Academic Performance Evaluation of Students (APES) is undertaken as an integral component of the Teaching Learning Evaluation Review (TLER) Model Environment.

5. Quantization, Contraction-Expansion Mappings, Information Loss/Corruption

Before we take up the task of proposing an alternative system design, let us first probe deeper into the causes for the above observed system behavior. The measurement scale for raw score is usually of a better (smaller) precision than that of the *Quality Points Scale* for the letter grades. *The Scale Precision is generally defined as a fraction of (relative to) the Scale Range.* However, in the case of *discrete scales* (with a discrete set of points, as the only valid measurements possible, rather than having an interval range) the scale-precision may equivalently be expressed as the reciprocal of the number that is *one less than the cardinality* of the set of valid measurement values that such a discrete scale accepts. The set of possible raw score values that a teacher may use in actual evaluation, say, using a scale range of 0-100 and a precision of 0.01 or 1-in-100, thus giving rise to 101 distinctly different possible values for a raw score, or equivalently, 101 distinct points as markings on the raw scores scale, labeled by the integers from 0 to 100, is of cardinality 101. The set of acceptable or valid letter grades (each associated with a quality point value used for computation of GPA) is usually of a small cardinality, like 5, or 6, or 11, or 16.

Any mapping scheme that one designs, to convert *Raw Score* to *Letter Grade*, would necessarily be a *Contraction Mapping Scheme*; that is, the "range" of the mapping is of smaller size than the "domain" of the mapping. Further, any contraction mapping between discrete sets, is necessarily a many-to-one mapping, thus resulting in *Information Loss/Corruption*.

Even imagining that we are considering the measurement scale to be continuous within its range, usually the mapping schemes that are prevalent, as is the one in the typical LGS system model explained above, do incorporate some *Quantization* of the raw scores scale. In other words, the raw scores scale is divided into some appropriate fixed number of *Quantization Interval Domains*, each of which is *mapped onto a single point/value in the discrete scale* of the set of letter grades; thus resulting in an effective *non-selectivity or non-differentiability* within each of these quantization

interval domains. The fact that, for later computations, each letter grade is associated with a point/value on a seemingly continuous scale of the quality points, is yet another mysterious system characteristic feature, that seems to have been incorporated in the overall system design. Anyway, it is essential to note here, that every instance of quantization is invariably associated with the occurrence of sharp *Quantization Interval Domain Boundaries*, which are points of *discontinuities* in the system behavior.

From the point of view of quantization, the Raw Scores Scale can also be considered as a quantized scale, in fact with a smaller scale quantum interval domain size. *The Scale Quantum Unit Interval Domain Size (SQUIDS) is defined to be the smallest discernible measurement interval that a scale can measure. The Scale Quantum Unit Interval Domain Size (SQUIDS) is in fact the Scale Precision.* It is called the Relative Precision, if expressed as a fraction of, or relative to the Scale Range; otherwise, it is the Absolute Precision. A contraction mapping as said above, from the raw scores scale to the *Quality Points Scale* (set of letter grades) necessarily results in the an effective lumping together of some of the neighboring quantum unit interval domains into an agglomeration, thus enlarging/widening the effective Scale Quantum Unit Interval Domain Size, expressed as a fraction of the corresponding scale range. So, the effect of contraction mapping in the case of discrete scales, is exactly the same as that of quantization on continuous scales; and that is, not only to worsen the scale precision, but also to introduce *discontinuities in the system characteristics*. This is how a situation arises wherein not only some crucial information is lost, but also a manifestation of the chaotic, highly complex (counter-intuitive), seemingly unpredictable system behavior, is observed, effectively introducing chaotic system biased information corruption as well.

It is to be noted that in any general Information System, when one passes through a transformation among the several alternative information representation schemes, there is a possibility of *Information Loss/Corruption* associated with such instances of *Quantization* and/or *Contraction Mapping*, followed by an implicit *Expansion Mapping* (corresponding to the mysterious presumption of a significantly higher degree of precision in the resultant combined value; the GPA and CGPA here). The only case that would not result in Information Loss/Corruption, is the one with the *Raw Scores Scale* and the *Quality Points Scale* to have exactly the same Range and Precision, thus having a *one-to-one onto mapping* transformation, for conversion between them. But, of course, this is far from reality, as observed in actual practice. From the point of view of Computational Systems Theory, the above observed *Information Loss/Corruption*, as well as the *chaotic and counter-intuitive system behavior*, are caused by the *drastically poor scale precision* used in the *intermediate representations*, that is, by the set of letter grades, or equivalently, the *Quality Points Scale*.

In the above example, the scale precision associated with the raw scores scale is 0.01 or 1-in-100. All the intermediate computations are performed on a very crude/rough/coarse scale of very poor precision; associated with each of the several discrete decisions regarding the specific choice of the letter grade to be assigned to each of the students in each of the subjects, from among the available set of assignable letter grades, that is of cardinality 6, thus having a scale precision of 0.20 or 1-in-5. However, the precision expected at the end, in the GPA and/or the CGPA as reported in (or, otherwise invariably determined, from) the transcripts/grade-reports is about 0.002 or 0.01-in-5.00 (that is, an accuracy of upto the second decimal place in the GPA and/or the CGPA figure,

on a scale range of 0-5). This significantly higher precision level somehow mysteriously presumed to have been achieved in the final figures of GPA or CGPA etc, is implied from the Expansion Mapping used in their computations as the weighted average of the Quality Points corresponding to the letter grades. This is exactly like the case of using single precision arithmetic for all intermediate computations wherein the initial input data is in double precision and the final result is also expected to be in double precision. The lost information cannot be recovered, but only get corrupted by this computational approach! Now, imagine how it is possible at all, to expect *unbiased, fair, and reliable* representation and transfer of information by/through such a system!

Also, it is to be noted here that the use of Course Credits/Units (CCU) as multiplicative weighting factors to the quality-points (as well as the quality-points-scale range) associated with the letter grades, cannot result in any improvement in the scale precision, although the range does get widened/enlarged, since the scale quantum interval domain size also gets widened/enlarged by exactly the very same multiplicative factor.

**6. Measurement-Scale: Type, Range, Precision,
Scale Quantum-Unit Interval-Domain Size (SQUIDS)
And
Scale Fixed Points**

Any rational or systematic evaluation is possible only through appropriate measurements of the characteristic properties or criteria for such evaluation. Any measurement and therefore any evaluation is possible only through comparisons; and even the very purpose of evaluation is for comparison among the entities evaluated using the criteria measured. If it were not for comparisons, no measurement is possible, nor any evaluation required.

Any measurement system must incorporate an appropriately designed scale for the measurements. The nature of the entities being characterized through such measurements, and also the kind of the specific attribute that is being measured, determine the type of the measurement scale that may be appropriate for the purpose. These details lead to the establishment of some well-accepted standards for such measurements, with appropriate measurement scales defined. These measurement standards and the associated scales for measurements may be classified into several broad categories. First consider a *nominal scale*, that is used for characterization of nominal entities that require some kind of distinct labeling of the different possible attribute values, with no specific relational order among them. For example, the color of some object may be white whereas that of another may be black, and this attribute (maybe along with other possible attributes, in combination) may be used in the characterization and/or the distinct identification of one object from another. So, a finite *discrete set* of colors from a possible collection of identifiable colors in a palette (based on the technology available) may serve the purpose.

Very soon we end up in a situation which forces us to improve our measurement system, if we think of two or more objects that are somewhat gray, somewhere between pure white and pure black. One may extend the finite discrete set of colors in the palette, but that process may result in a very

unwieldy measurement scale to work with in practice. So, a reasonable relational operator is imposed on the attributes, so that there is a structure for the set of valid values that an attribute can take. In the case of colors, there exists a well established system to characterize the visual appearance of an object from the point of view of the measurements associated with the color, hue, brightness, etc. of the light received from the object (reflected or deflected or emitted, etc) described through appropriate parameter values (e.g., the RGB co-ordinates and the intensity/brightness, or even a detailed Spectral Distribution data). A measurement scale when used for relative comparisons among values that an attribute takes on the scale, corresponding to two or more objects, can be considered to be a *relational scale*. For example, one object may be of relatively lighter blue color, when compared with a second object that is of darker blue color, and yet another object that is of deep blue color. Here, it may not be meaningful to take the numerical differences in the attribute values, and say for example, that the first object is as much lighter in its blue color when compared to the second object, as the third object is darker in its blue color when compared with the second object. In other words, although the measurement scale admits relational operators (comparisons), the numerical differences, or the interval sizes, or the distances, among the attribute values may not be meaningfully defined!

Now consider, as an example, the Centigrade (or even a Fahrenheit) scale for temperature measurement, using which the temperatures of two or more objects can be compared, in the sense than one could be either more or less than another by some extent/degree, and such differences in extent/degree can be well characterized using the same measurement scale. A measurement scale that allows the direct measurement of (or a later computation) of the extent/degree of the differences in attribute values (through some well defined concept of interval size, or that of a distance measure, or simply the numerical difference) can be considered to be a *relative interval scale*.

In an interval scale, although the interval sizes, or the distances, or simply the numerical differences in the attribute values, are well defined, there may not exist any absolute fixed point in such a measurement scale. Now, if we consider the Absolute Temperature Scale (degree-Kelvin) it turns out that there exists what can be considered as an *Absolute Fixed Point* in such a scale. The very measurement in such a scale is a direct and explicit comparison of an attribute value relative to that absolute fixed point. Such a measurement scale, with some (one or sometimes more) absolute fixed points, is considered as an *Absolute Scale*. The attribute values measured in an absolute scale may allow for (1) meaningful relative comparisons among attribute values, (2) meaningful concept of the interval sizes, or a distance measure, or simply the numerical differences in the attribute values, (3) meaningful computations of the numerical ratios of the interval sizes, (4) meaningful concepts associated with the distances, w.r.t the absolute fixed point/points, and their ratios.

A complete description of a measurement scale includes the specification of the *Scale Range* (cardinality of the set of possible values, in the case of a finite discrete scale) and the *Scale Precision or the Scale Quantum Unit Interval Domain Size (SQUIDS)*, and also the *Scale Fixed Points*, if any.

At this point, it is very useful to note how the incorporation of an appropriately chosen design modification in a system, although seemingly

simplistic, can in fact result in a significant enhancement in the system characteristics. Specifically, consider the incorporation of a negative one-fourth (-0.25) mark/score for every wrong answer, in a written aptitude test with 100 questions and each correct answer carrying one (+1) mark/score. It is very easy to see that the scale range gets enlarged/widened; that is, from -25 to +100 now, as against the earlier range of 0-100. The scale quantum unit interval domain size gets reduced to 0.25 now, as against 1.00 earlier. The result is a five-fold improvement in the scale precision, that is, 0.002 or 0.25-in-125 or 1-in-500 now, as against 0.01 or 1-in-100 earlier.

A finer scale, with a better precision allows the teacher to incorporate finer distinctions in the various possible performance levels encountered during the evaluation process, and does not force the teacher to be restricted by the limitations of a coarse/crude/rough scale with poorer precision. A coarser scale (poorer precision) gives rise to undue magnification of the differences between the possible measurements that happen to be located on either side just across the border between adjacent scale quantum unit interval domains. Of course, if a teacher does not need the precision provided by the measurement scale incorporated in the system design; a suitable mapping or superimposition of an appropriately chosen coarser scale onto the finer scale (incorporated in the system design) can in fact be performed with neither any Information Loss/Corruption nor any undesirable system behavior like the ones observed and analyzed here. For example, a teacher may decide to use only the eleven numbers 0, 10, 20, . . . 100; instead of the 101 possibilities, even though provided with a percentage scale with the scale range 0-100 and scale precision of 1-in-100. It is generally advisable however, for each teacher to effectively utilize the full representational capability, and the associated selectivity, that is made available through the measurement system design.

From the analyses presented above, it is clear that the most fundamental basis in the design of any measurement scale is in fact, the *Scale Quantum Unit Interval Domain Size (SQUIDS)*, which defines the *Scale Precision*, which together with the *Scale Range* and the *Scale Fixed Points (if any)* provides a complete description for the measurement scale.

The undesirable system behavior explained earlier, are the result of a poor system design, specifically in terms of the implicit changes both in the *type of the scale* as well as the *scale range* and particularly the *scale precision* or the *scale quantum unit interval domain size*, as incorporated into the system design at the different stages of information flow through the system. Or, rather there is a *possible ambiguity* as to the appropriate application of the very concept of scale quantum unit interval domain size, and that of scale precision (only the concept of scale range seems to have been explicitly incorporated) even if implicit in the system design.

Specifically, in the first stage the teacher is given an almost complete freedom as to the *definition of a measurement scale* to be used for raw scores; the system design does not address this issue at all. The system design is ambiguous in the second stage, wherein a mapping scheme or transformation is to be defined to convert the raw scores to the letter grades defined as a finite discrete set of relatively small cardinality (on which is possibly defined an anti-symmetric transitive binary relation, "better-than"). This results in a drastic deterioration of the scale precision (defined as the scale quantum unit interval domain size). In the third stage, there is an almost immediate usage of a well defined one-to-one onto mapping to the *Quality Points Scale*, as defined and specified by the

system design, thus maintaining the same low/poor precision. In the fourth stage, the Course Credits/Units (CCU) are used as multiplicative weighting factors in the computation of a combined aggregate or overall score (namely, the GPA) in the *Quality Points Scale*, with an expectation of impossibly high precision level as mentioned earlier.

In addition to these changes in the nature of measurement scale used, the use of the Course Credits/Units (CCU) as multiplicative weighting factors in the fourth stage, that is the computation of the overall score (namely, the GPA) is again not a good systems design. For example, consider a situation wherein a teacher evaluates two students, the first one being at 90% level, and the second one being at 89% level, using a percentage scale, their difference being just equal to the scale quantum unit interval domain size (SQUIDS). When multiplied by the CCU value of 4, for example, these numbers become 360 and 356 respectively. Now, let us ask the question, why shouldn't the teacher be allowed to use the enhanced scale range directly for the purpose of the evaluation? It might so happen that, if the teacher were to be given this facility of an enhanced scale range right at the initial stage of evaluation itself, with the same fixed value of the scale quantum unit interval domain size (SQUIDS), that is, the absolute scale precision, then, the actual evaluations could possibly as well have been anywhere in the range 358-361 for the first student, and in the range 354-357 for the second student, out of 400 points. This design feature would have given a better measurement scale on the hands of the teacher right at the initial step of evaluation, and thus improve the effective selectivity or the effective differentiability of the entire evaluation system. From this point of view, the incorporation of multiplicative weighting factors in the computation of the overall score, is in fact *analogous to a delayed patching up for a lost opportunity without any real benefit.*

It is to be noted here that, in situations where numerical multiplication and/or division operations are carried out, the concept of *relative precision* expressed as a fraction of (relative to) the scale range, is relevant; whereas, in situations where numerical addition and/or subtraction operations are carried out, the concept of *absolute precision* as the scale quantum unit interval domain size (SQUIDS), is relevant.

It is useful to note here that, a typical raw scores scale used in many of the international competitive examinations like GRE, TOEFL, GMAT, etc. can be seen to have a range of at least 1000, with a precision appropriate for determining the corresponding percentile rankings represented with the second decimal place accuracy, since the number of candidates appearing for such tests may run into a four-digit number. Imagine, what would be the consequences (in terms of the effective selectivity) if these competitive examinations were designed to use a five-point or even a sixteen-point grading system to declare their results!

In many situations, it is not only very essential that the entire evaluation and measurement system is designed to be *unbiased, fair, and reliable*; but also it is desirable to incorporate appropriate *elimination and selection* mechanisms, with *screening and filtering* steps, so that at the end, the evaluation and measurement system has a high degree of *selectivity or differentiability or distinguishability* among the various possible candidate performance levels.

7. Student Academic Performance Evaluation System (SAPES)

Now, let us apply the above ideas, in the context of the Students Academic Performance Evaluation System (SAPES) that we like to design. When a teacher designs a syllabus for a course, there is an expectation in the mind of the teacher, regarding the extent or degree of academic proficiency that a student is expected to achieve, by studying through that specific course. Whether such an expectation is made explicit or otherwise remains implicit, is immaterial at this point for our analysis and design. The extent or degree of academic proficiency that a student achieves through the learning process, is evaluated or measured by the teacher as an evaluator. In order to make these evaluations as objective as is realistically possible, the academic proficiency achieved is in fact measured through the evaluation or measurement of the academic performance of a student in a (seemingly continuous) sequence of tests or examinations or other evaluation schemes. The performance level of each of the students, in each of such tests is evaluated by the teacher, and the combined result of such evaluations, is considered to be an acceptable measure of the academic proficiency achieved by the student in that course, during a particular academic term.

Even without considering the situation for comparisons among the various possible student performance levels, in any subject, it is possible for a teacher to imagine having an absolute measurement scale, with two fixed points, namely the *NULL Performance Point*, and the *FULL Performance Point*; usually corresponding to the 0% point and the 100% point on a percentage scale. Both these fixed points are usually very well identifiable, for a teacher, and is directly defined in terms of the extent of academic proficiency that a student is expected to achieve by the study of that course. The tests/exams/etc can be designed carefully, to be able to measure, as well as to distinguish the various possible performance levels of the students. One need not unnecessarily bring in the complications of possible non-linearities in the scale for this purpose. The non-linearities occur (if at all) mainly because of poorly specified syllabus or poorly designed tests; and should not be considered as intrinsic in the *design* of the *measurement system structure and mechanism*. It is clear that any well designed test (or a sequence of tests, or some other evaluation scheme) must incorporate an appropriately selected set of questions such that through such an evaluation, the true level of academic proficiency of every student in the class could be measured in an *unbiased, fair and reliable measurement scheme*. For this purpose, it is necessary that the test questions must evenly span the breadth and depth of the entire syllabus of the particular course. Also, it is certainly desirable in that process of evaluation, to be able to clearly distinguish the various possible performance levels, without much ambiguity. It is certain that a poorly designed test may end up being a curse to an academically proficient student, one who may not get an opportunity to distinguish oneself through one's performance (in such a test) from every other student; whereas it can even be a boon to another one who is in fact not so proficient, but rather happens to be just simply lucky in this particular instance. No system can be made completely fool-proof in this respect, except by the incorporation of appropriate checks and balances, in the system policies and procedures, or the rules and regulations; and we will not get into these aspects, here in this paper.

8. Proposed Design for SAPES

It is proposed that the measurement scale must be given the central focus in the design of the Students Academic Performance Evaluation System (SAPES). *The Scale Quantum Unit Interval Domain Size (SQUIDS) is fixed to be equal to UNITY*, and is used as the most fundamental basis, as a single, universally common, unique standard unit of measurement, in the design of a unique SAPES measurement scale, which I call as the *SAPES Scale*. The Scale Range can be represented by identifying the Scale Minimum Point (SminiP) and the Scale Maximum Point (SmaxiP), which usually (but not necessarily, in general, as we shall see later) may correspond to the NULL Point and the FULL Point respectively.

The *Student Academic Performance (SAP) Scale Range* must correspond to the entire spectrum of possible academic performance levels, between the Scale Minimum Point (SminiP) and the Scale Maximum Point (SmaxiP). Usually, the two well defined *Scale Fixed Points*, the *NULL Performance Point* and the *FULL Performance Point* define the effective SAP Scale Range.

The SAP Scale can be so designed for each of the different courses, that the Null Performance Point is common for all these different SAP Scales. This has an intuitive appeal as well, since it takes the same (zero or nil) extent of achievement of academic proficiency in order to perform at the NULL-Performance level, irrespective of whichever be the specific course being considered. This universally common *Null Performance Point* is the common zero point for every SAP Scale, irrespective of the course/subject, and is therefore called the *SAP Absolute Zero Point*, in the *SAPES Measurement Scale System*.

With this common SAP absolute zero point, and with the SQUIDS as a common fundamental basis for the unit of SAP scale measurements; it is clear that the only variations possible in the different SAP scales corresponding to the different courses, is the location of the FULL Performance Point, and the SAP Scale Range for the particular course. With this design, it is possible to have a *single, universally common, unique Standard SAP Scale*, which we will refer to as the *SAPES Scale*. That is, the SAP scales corresponding to the various different courses are in fact identical to one another, except for the location of the FULL Performance Point and the SAP Scale Range corresponding to each of these courses, on the *SAPES Scale*.

The location of the FULL Performance Point corresponding to a specific course, on the SAPES Scale; or equivalently, the number of Scale Quantum Unit Interval Domains, between the two fixed points on the SAP scale corresponding to that specific course; is determined by the number of *SAP Credits (SAPC)* for that course, as expressed in the above mentioned "SQUIDS" unit of measurement (*one SAPC is indicated in the SAPES Scale by one SQUIDS*).

For the sake of easy understanding, the SAP Credits (SAPC) for a course can be considered to be somewhat related to the currently prevalent concept of the Course Credits/Units (CCU) in the existing system. Although the two are numerically not the same, one can in fact draw a simple relationship between the two, by trying to match the expected/desired behavior of the existing APES system, to what can in fact be actually observed with the new system design. With this in mind, one can consider the SAP Credits for a course to be simply *proportional* to the number that refers to the Course

Credits/Units (CCU). For example, suppose that one *Course Credit/Unit* (CCU) is assigned 100 *Credit Unit Score Points* (CUSP), and that one CUSP is assigned one SAPC. Then, a 4-credit course that is assigned with 4 CCU, can now be considered to carry 400 CUSP or 400 SAPC). The FULL Performance Point for this course would be a point on the SAPES Scale that is marked as 400 SQUIDS (or equivalently, 400 CUSP, or 400 SAPC). On these same lines of reasoning, the FULL Performance Point for a course with 3 CCU, that carries only 300 CUSP or 300 SAPC, would be a point on the SAPES Scale that is marked as 300 SQUIDS (or equivalently, 300 CUSP, or 300 SAPC), etc. That is, the course/subject specific FPP is identified by, its CUSP value or equivalently its SAPC value. It is apparent that in the SAPES Environment, each course/subject must be assigned a SAPC value or equivalently a CUSP value, in order to be accounted for consideration towards the award of a SAPS score for the student who registers oneself for that course for SAP Credit. The *Scale Precision* (one SQUIDS) on our SAPES-Scale which is equal to one SAPC, now corresponds to a CUSP value of *unity*, which is equivalent to a CCU value of *one-hundredth of unity*.

It may be useful to note here that the above mentioned Absolute Zero Point remains as the same common fixed point for even a SAP Scale that may need to be designed to incorporate even negative values for possible SAP raw score measurements. For example, in the case of the written aptitude test scenario mentioned in an earlier section of this paper, one may design a special SAP scale with the Scale Minimum Point (SminiP) as a point on the SAPES Scale that is marked as -100 SQUIDS (-25 Marks), corresponding to the lowest possible performance level; and the Scale Maximum Point (SmaxiP) as a point on the SAPES Scale that is marked as +400 SQUIDS (+100 Marks), corresponding to the highest possible performance level; with the SAPES Absolute Zero Point as the universally common NULL Performance Point. The SQUIDS value for such a SAP Scale would then correspond to 0.25 Marks as usually awarded. That is, each correct answer gets a positive score of four SQUIDS (+1 Mark), and each wrong answer gets a negative score of one SQUIDS (-1/4 Mark) in the SAPES Measurement Scale. So, the value of any score as usually awarded in Marks can be easily and directly converted to the corresponding SQUIDS value on the SAPES Measurement Scale, by just multiplying by a factor of four, since one Mark corresponds to four SQUIDS (for example, refer to conversion of Absolute Temperature figures expressed in degrees Rankine and degrees Kelvin by the appropriate multiplication factor; unlike in the case of conversion between degrees Centigrade and degrees Fahrenheit). Thus, the SAPES Measurement Scale provides a universally common structural basis as a rational and systematic framework to combine/compare the scores obtained in (pertaining to) various distinctly different environments (like, for example, the scores obtained in Written Aptitude Test, along with that contained in the Transcripts or Grade Reports, etc).

Every teacher is expected to assign a valid *SAP Raw Score* simply called as the *SAP-Score* (SAPS), for each of the students, in each of the courses that a student "registers for SAP credit", which has to be a positive integer in the range from zero to a maximum possible value given by an integer number equal to the SAPC or the CUSP for the specific course. The individual teacher may be given the freedom as to how to arrive at a valid SAP-Score (SAPS), which is an appropriate, unbiased, fair, and reliable measure of the student academic performance; possibly using a sequence of tests, etc; about which, some general guidelines may be offered by the department or school or college or university. No intermediate conversions or transformations are to be performed, which would affect the information content represented by these

SAP Raw Scores. These SAPS scores associated with the various courses (corresponding to a semester, or an academic term/session, or the entire degree programme) are *just simply* added together to compute a sum (tSAPS) total of all the SAPS scores, or even possibly a running cumulative if needed. Finally the *Student Academic Performance Score index (SAPSi)* is simply the ratio, $(tSAPS)/(tSAPC)$, or equivalently $(tSAPS)/(tCUSP)$. Here, tSAPC is the sum total of the SAP Credits (SAPC) or equivalently the Credit Unit Score Points (CUSP) of all the courses that are to be accounted for, as per the graduation requirements, and other applicable policies and procedures or rules and Regulations of the school/college/university. The Student Academic Performance Score index, SAPSi may either be reported as such, explicitly giving the two numbers tSAPS and tSAPC (or equivalently, tCUSP); or as a fraction or even as a percentage value representing that ratio. Similarly, the *cumulative Student Academic Performance index (cSAPSi)* of a student for the entire programme of study, can be computed. Some of these various figures may as well be expressed as a percentage; for example, SAPS expressed as a percent of SAPC (or equivalently, CUSP); tSAPS expressed as a percent of tSAPC (or equivalently, tCUSP); and/or cSAPS expressed as a percent of cSAPC (or equivalently, cCUSP); for convenience in communication, and possibly to check for the individual course requirements (like minimum for Pass, course Distinction, etc, if at all such things do matter). However, all the original evaluations or measurements, all the intermediate representations, and also all the computations, are to be performed conforming to the SAPES Scale, as explained, so as not to introduce any undesirable system characteristics.

Now, let us get into further explanations on the above proposed system design. Having a single unique Standard SAPES Scale, as the central system component, and having fixed the fundamental/basic standard unit of measurement to be the SQUIDS unit, in the design of that measurement scale, means that every teacher uses a mark/score of UNITY (and not any fractional mark/score), as the smallest quantity of mark/score for the purpose of any evaluation. In special situations requiring a non-unit value for the SQUIDS, like in the case of the written aptitude test mentioned earlier in this paper, appropriate conversion has to be performed as illustrated there, in order to express the SAP Score in the *Universally Common Unique Standard SAPES Absolute Scale* with the Absolute Scale Precision (SQUIDS) of UNITY and the universally common Scale Fixed Point at NULL Performance Point. This standardizes the entire world of Student Academic Performance Evaluation System (SAPES), into a manageable entity for the purpose of design, analysis as well as implementation, without in any way causing undue restrictions or any inconveniences to the teachers, or the students or any other personnel who is required to work with the system. Any ambiguities in system definition, as well as system complexities exhibited through counter-intuitive system behavior are all cleanly avoided by this approach.

9. Some General Comments on SAPES Environment

Now, let us address some of the possible questions that could be raised by the proponents of a "distribution based" evaluation scheme. It may be generally felt that the raw scores and therefore the SAPI index of student academic performance evaluations system described above, can be to a large extent possibly dependent on the relative leniency or strictness of the specific teachers or evaluators, except in special situations where the tests themselves are standardized, with objective multiple choice questions, each

associated with possibly one correct answer choice, or a best answer choice, to choose from the multiple choices provided, in which case the Scale Quantum Unit Interval Domain Size (SQUIDS) unit of measurement thereby gets defined clearly (although without a deliberate conscious design effort towards this specific aim of developing an unambiguous standard SQUIDS unit for measurement). However, the solution to such a problem of possible variability (dependency on the teachers' relative leniency levels) is certainly not to go for the "distribution based" evaluation scheme, which would exhibit an almost equal extent of variability and an explicit dependency on the distribution of the raw scores in a class, which again is directly dependent on the individual teacher's leniency level, etc. Also, the fact that all the undesirable system characteristics like the CSBAAD, CSBSED, and CSBRI still persists in this scheme. The only difference is that the location and extent/degree of the discontinuous system behavior, etc. would now be dependent on the Quantization Interval Domain Boundaries which get defined by each of the individual teachers at each instance as needed, and hence can be quite ad hoc in nature. So, in addition to all the above mentioned (and analyzed) undesirable system characteristics, this distribution based evaluation scheme also suffers from an added dimension of undesirability, due to this ad hoc system design, and therefore certainly requires one to look for a different alternative system design, anyway. A better solution to the problem of *teacher-to-teacher variability*, or rather the problem of *dependency on the teachers' leniency/strictness levels*, is in fact to provide appropriate information (in the transcripts/grade-reports or now we may call it the SAPS-Report) about such dependency. One can also possibly think of an extremely ideal situation, wherein the entire distribution of each of the SAPS scores is provided every time a specific SAPS score reported. But, this solution can of course become quite enormous to be of practical value. At least to start with, one can certainly think of incorporating in the overall SAPHES system, a policy to include some information regarding the *Class Academic Performance Statistics* (CAPS); specifically the five statistical parameters, namely, CAPSsize - the *Class Size* (number of students in the class), CAPSmini - the *Class Minimum SAPS score*, CAPSmaxi - the *Class Maximum SAPS score*, CAPSmean - the *Class Mean (Average) SAPS score*, and possibly CAPSmedi - the *Class Median SAPS score*; associated with each of the courses for which a SAPS score is reported. Such statistical data would certainly provide the right perspective to an interpretation of the specific SAPS scores. On the other hand, any ad hoc normalization or transformation based on some standard distribution curve will only result in Information Loss/Corruption, as mentioned earlier. How exactly these additional statistical data, associated with each of the SAPS scores are to be utilized in a systematic analysis of a set of student SAPS Reports (transcripts/grade-reports) is a subject for a possibly separate paper. The only point to be stressed here is that, each of the individual SAPS scores must not be tampered with any kind of normalizations or transformations or whatever, and must be left as such. Any further analyses on the data supplied, can always be carried out based on the actual need for such analyses.

A teacher who still decides/wishes to use fractional marks/scores in any evaluation or measurement, in this new SAPHES environment, may certainly be allowed to do so, as long as the SAPS score that is reported in fact corresponds to the SAP Scale for that specific course, with the SAPC value or equivalently the CUSP value, as the maximum possible value for any SAPS score thereof. The only comment here is that such a teacher may simply be over-exerting, with or without some well defined deliberate intent thereof. On the other hand, if the smallest chunk of mark/score that a teacher uses is

significantly larger than unity (that is, one SQUIDS unit) then it may seem to indicate that this teacher may not probably be effectively utilizing the available scale precision for the purpose of an unbiased fair and reliable evaluation/measurement of the students academic performance levels, contrary to the expectations of both the students community as well as the recruiters or employers, etc. Usually, this enables the department, or the school, or the college, or the university, to evolve appropriate policies and procedures, or rules and regulations, for establishing the necessary checks and balances in the overall SAPES environment. However, it is to be noted here that, there may be situations corresponding to some courses/subjects, wherein the very nature of the subject material and the possible methods available for measurements of a student's academic performance in such situations, impose certain constraints on the measurable precision levels thereof. It is important to note here that the available scale precision (SQUIDS) must be at least as much as or even better than that required in any/every measurement scenario that uses such a measurement scale. In other words, while using any measurement scale, the best achievable measurement precision cannot be better than that provided by the measurement scale. However, if situation demands, we may always be able to use the same measurement scale for actual measurements requiring relatively less precision than what is provided by the measurement scale.

Also, it is interesting to note that the SAPES measurement scale, and the system design in principle, does not necessarily prevent a teacher from assigning a SAPS score the numerical value of which is larger than that of the corresponding SAPC credits, and therefore may at the outset seem to be invalid as per the earlier discussion. However, even such exceptional situations, indicating that the academic performance level of that particular student in that specific course has been evaluated to be beyond the expected standard FULL Performance level, may be allowed. Here, and in such similar situations, it is of course desirable to require certain appropriate specific approvals from the department and/or the school and/or the college and/or the university, so as to provide appropriate checks and balances, in terms of policies and procedures, or rules and regulations, incorporated into the overall system environment. It is quite a simple matter to implement such a possible policy/regulation on the possible validity of the SAPS score, if so required, to be within some appropriately specified limits identified by the Scale Minimum Point (SminiP) and the Scale Maximum Point (SmaxiP).

As a final note, let me add here that the SAPES Environment may very well be used even if someone decides to go with the idea of "relative evaluation" or "distribution based evaluation", that is equivalent to what has been usually referred to as "grading on the curve" in the existing LGS system. The availability of an absolute scale of measurement can certainly be made use of even when a situation may in fact require relative evaluation, or even if a nonlinear mapping is required. However, I believe that the use of any nonlinear scale for measurement may introduce unnecessary complexity or may even cause loss of information, and therefore to be considered as undesirable, especially when such non-linearity cannot be universally justified. The concept of distribution based evaluation, can introduce significant non-linearities, and may possibly be associated with the uncertainties in the parameters of such distribution, thus introducing another dimension of complexity, in addition to causing information loss or rather an irrecoverable information mutation, in the system design.

* * *